Nucleic Acids Research, 1999, Vol. 27, No. 1 29–34

KEGG: Kyoto Encyclopedia of Genes and Genomes

Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono and Minoru Kanehisa*

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Received September 8, 1998; Revised September 22, 1998; Accepted October 14, 1998

ABSTRACT

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledge base for systematic analysis of gene functions in terms of the networks of genes and molecules. The major component of KEGG is the PATHWAY database that consists of graphical diagrams of biochemical pathways including most of the known metabolic pathways and some of the known regulatory pathways. The pathway information is also represented by the ortholog group tables summarizing orthologous and paralogous gene groups among different organisms. KEGG maintains the GENES database for the gene catalogs of all organisms with complete genomes and selected organisms with partial genomes, which are continuously re-annotated, as well as the LIGAND database for chemical compounds and enzymes. Each gene catalog is associated with the graphical genome map for chromosomal locations that is represented by Java applet. In addition to the data collection efforts, KEGG develops and provides various computational tools, such as for reconstructing biochemical pathways from the complete genome sequence and for predicting gene regulatory networks from the gene expression profiles. The KEGG databases are daily updated and made freely available (http://www.genome.ad.jp/kegg/).

INTRODUCTION

The progress in structural genomics will soon uncover the complete genome sequences for hundreds and thousands of organisms. However, roughly a half of the genes identified in every genome thus far sequenced still remain unknown in terms of their biological functions. New experimental and informatics technologies in functional genomics are urgently required for systematic identification of gene functions. Kyoto Encyclopedia of Genes and Genomes (KEGG) is an effort to computerize the current knowledge of biochemical pathways and other types of molecular interactions that can be used as reference for systematic interpretation of sequence data (1). At the same time, KEGG attempts to standardize the functional annotation of genes and proteins, and maintain gene catalogs for all complete genomes and some partial genomes including mouse and human.

The basic concepts of KEGG (1) and underlying informatics technologies (2,3) have already been published. KEGG is tightly integrated with the LIGAND chemical database for enzyme reactions (4,5) as well as with most of the major molecular biology databases by the DBGET/LinkDB system (6) under the Japanese GenomeNet service (7). The database organization efforts require extensive analyses of completely sequenced genomes, as exemplified by the analyses of metabolic pathways (8) and ABC transport systems (9). In this article, we describe the current status of the KEGG databases and discuss the use of KEGG for functional genomics.

OBJECTIVES OF KEGG

In May 1995, we initiated the KEGG project under the Human Genome Program of the Ministry of Education, Science, Sports and Culture in Japan. We wish to automate human reasoning steps for interpreting biological meaning encoded in the sequence data. We consider the problem of predicting gene functions as a process of reconstructing a functioning biological system from the complete set of genes and gene products. Thus, it is critical to understand how genes and molecules are networked to form a biological system. Specifically, the objectives of KEGG are summarized in the following four points.

First, KEGG aims at computerizing the current knowledge of genetics, biochemistry, and molecular and cellular biology in terms of the pathway of interacting molecules or genes. The KEGG pathway database contains the information of how molecules or genes are networked, which is complementary to most of the existing molecular biology databases that contain the information of individual molecules or individual genes. During the first two years of the KEGG project we focused on the metabolic pathways, but since July 1997 we have also been collecting a large body of knowledge in the regulatory aspects of cellular functions.

Second, KEGG maintains the gene catalogs for all organisms with completely sequenced genomes and selected organisms with partial genomes. Because the criteria of interpreting sequence similarity are different for different authors, the quality of gene function annotations varies significantly in GenBank (10). KEGG's gene catalogs are intended to provide consistent and standardized annotations by linking individual genes to components of the KEGG biochemical pathways.

Third, KEGG maintains the catalog of chemical elements, compounds, and other substances in living cells as the LIGAND

*To whom correspondence should be addressed. Tel: +81 774 38 3270; Fax: +81 774 38 3269; Email: kanehisa@kuicr.kyoto-u.ac.jp

Table 1. The summary of KEGG release 8.0 (October 1998)

Database / auxiliary data	Number of entries	Contents	Direct links	Additional links by DBGET/LinkDB*
PATHWAY	2,085 entries	GIF image map showing biochemical pathways	GENES, ENZYME, COMPOUND, other PATHWAY	GenBank, EMBL, SWISS-PROT, PIR, PRF, PDB, FDBSTE, IFO, TRANSFAC, PROSITE, OMIM, PMD, AAindex, LITDB, Medline, similar sequences
GENES	79,891 entries in 17 complete genomes and 12 partial genomes	Nomenclature, classification, codon frequency, amino acid sequence, nucleotide sequence	PATHWAY, Gene catalog, Genome map, ENZYME, SWISS-PROT	
LIGAND/COMPOUND	5,386 entries	Nomenclature, chemical formula, molecular structure, CAS number	PATHWAY, ENZYME	
LIGANDENZYME	3,391 entries	EC number, nomenclature, classification, reaction formula, substrate, product, inhibitor, effector, cofactor	COMPOUND, PATHWAY, GENES, OMIM, PROSITE, PDB, EnPAS+, WIT, UM-BBD, BRENDA, SCOP	
Genome map	17 complete genomes and 1 partial genome	Java applet for searching and browsing genes in a genome map	GENES, PATHWAY	
Comparative genome map	17 x 17 complete genome comparisons	Java applet for comparing genes in two genome maps	GENES	
Ortholog group table	4,772 genes in 51 tables	HTML table for functionally related genes, orthologs and paralogs, and operon structure	PATHWAY, Genome map, GENES	
Gene catalog	43 catalogs	Classification of genes (hierarchical text data)	PATHWAY, GENES, ENZYME, SWISS-PROT	
Molecular catalog	8 catalogs	Classification of proteins, RNAs, and compounds	GENES, ENZYME, COMPOUND, SWISS-PROT	

*Additional links can be obtained by clicking on the entry name of each entry; in the case of a graphical pathway map, by clicking on the title box.

database (4,5) and they are again linked to pathway components. This is based on our view that both the genetic information encoded in the genome and the chemical information encoded in the cell are required for understanding cellular functions.

Fourth, in addition to the database efforts, KEGG aims at providing new informatics technologies that combine genomic and functional information towards predicting biological systems and designing further experiments. For example, the KEGG reference pathways can be used to uncover molecular interactions and pathways that underlie gene expression profiles obtained by microarray experiments (11).

THE KEGG DATABASES

All the data in KEGG can be accessed from the KEGG table of contents page (http://www.genome.ad.jp/kegg/kegg2.html), which is divided into two broad categories: the pathway (functional) information and the genomic (structural) information. The user may enter the KEGG system top-down starting from the pathway information or bottom-up starting from the genomic information. In either case it should be easy to navigate through KEGG since all the information is highly integrated. The KEGG data are also linked with many other molecular biology databases by the DBGET/LinkDB system. A summary of KEGG data contents is shown in Table 1.

PATHWAY database

The KEGG/PATHWAY database is a collection of graphical diagrams (pathway maps) for the biochemical pathways. Table 2 shows the current list of the KEGG biochemical pathways, which is highly biased toward the metabolism. There are about 90 reference maps for the metabolic pathways that are manually drawn and continuously updated according to biochemical evidence. The organism-specific pathway maps are then automatically generated by matching the EC numbers in the gene catalogs and in the reference maps. The maps for the regulatory pathways are drawn separately for each organism, since they are too divergent to be represented in a single reference map.

Figure 1 shows an example of the KEGG metabolic pathway map; citrate (TCA) cycle for *Haemophilus influenzae* and *Helicobacter pylori*. A box is an enzyme (gene product) with the EC number inside and a circle is a metabolic compound. They are both clickable objects to retrieve detailed molecular information. The enzymes (boxes) whose genes are identified in the genome are colored green (shaded in Fig. 1) by the process of matching the gene catalog and the reference pathway according to the EC numbers. It is interesting to note that the two organisms seem to

Table 2. The KEGG biochemical pathways (release 8.0, October 1998)

Metabolism

Carbohydrate Metabolism
Glycolysis / Gluconeogenesis
Citrate cycle (TCA cycle)
Pentose phosphate cycle
Pentose and glucuronate interconversions
Fructose and mannose metabolism
Galactose metabolism
Ascorbate and aldarate metabolism
Pyruvate metabolism
Glyoxylate and dicarboxylate metabolism
Propanoate metabolism
Butanoate metabolism
C5-Branched dibasic acid metabolism
Energy Metabolism
Oxidative phosphorylation
Methane metabolism
Carbon fixation
Reductive carboxylate cycle (CO2 fixation)
Nitrogen metabolism
Sulfur metabolism
Lipid Metabolism
Fatty acid biosynthesis (path 1)
Fatty acid biosynthesis (path 2)
Fatty acid metabolism
Synthesis and degradation of ketone bodies
Sterol biosynthesis
Bile acid biosynthesis
C21-Steroid hormone metabolism
Androgen and estrogen metabolism
Nucleotide Metabolism
Purine metabolism
Pyrimidine metabolism
Nucleotide sugars metabolism
Aminosugars metabolism
Amino Acid Metabolism
Glutamate metabolism
Alanine and aspartate metabolism
Glycine, serine and threonine metabolism
Methionine metabolism
Cysteine metabolism
Valine, leucine and isoleucine degradation
Valine, leucine and isoleucine biosynthesis
Lysine biosynthesis
Lysine degradation
Arginine and proline metabolism
Histidine metabolism
Tyrosine metabolism
Phenylalanine metabolism
Tryptophan metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis
Urea cycle and metabolism of amino groups
Metabolism of Other Amino Acids
beta-Alanine metabolism
Taurine and hypotaurine metabolism
Aminophosphonate metabolism
Selenoamino acid metabolism
Cyanoamino acid metabolism
D-Glutamine and D-glutamate metabolism
D-Arginine and D-ornithine metabolism
D-Alanine metabolism
Glutathione metabolism
Alkaloid biosynthesis I
Alkaloid biosynthesis II
Metabolism of Complex Carbohydrates
Starch and sucrose metabolism
Biosynthesis and degradation of glycoproteins
Peptidoglycan biosynthesis
Metabolism of Complex Lipids
Glycerolipid metabolism
Inositol phosphate metabolism
Sphingolipid biosynthesis
Phospholipid degradation
Sphingoglycolipid metabolism
Fatty alcohols and isoketones metabolism
Metabolism of Cofactors, Vitamins, and Other Substances
Thiamine metabolism
Riboflavin metabolism
Vitamin B6 metabolism
Nicotinate and nicotinamide metabolism
Pantothenate and CoA biosynthesis
Biotin metabolism
Folate biosynthesis
One carbon pool by folate
Retinol metabolism
Porphyrin and chlorophyll metabolism
Terpenoid biosynthesis
Xenobiotics metabolism
Ubiquinone biosynthesis
Flavonoids, stilbene and lignin biosynthesis
Metabolism of Macromolecules
mRNA biosynthesis (bacteria)
mRNA biosynthesis (eukaryotes)
tRNA and rRNA biosynthesis (bacteria)
tRNA biosynthesis (eukaryotes)
rRNA biosynthesis (eukaryotes)
Protein biosynthesis (bacteria)
Protein biosynthesis (eukaryotes)
Aminoacyl-tRNA biosynthesis

Cell Processes

Membrane Transport
ATP-binding cassette (ABC) transport system
Phosphotransferase system (PTS)
Signal Transduction
Two-component system
Bacterial chemotaxis
MAPK signalling pathway
Second messenger signalling pathway
Ligand-Receptor Interaction
G protein coupled receptors
Cytokine receptors

Cell Organization

Molecular Assembly
Ribosome assembly
Flagellar assembly
F1FO ATP synthetase

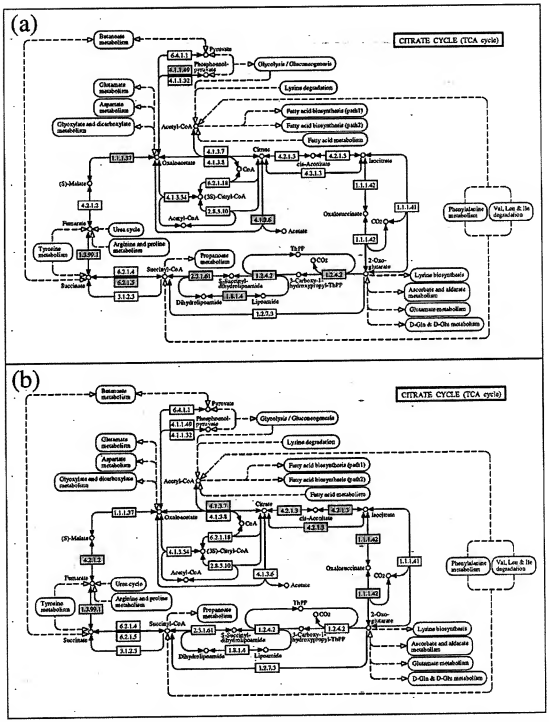

Figure 1. The KEGG pathway map of citrate (TCA) cycle for (**a**) *Haemophilus influenzae* and (**b**) *Helicobacter pylori*. A rectangle and a circle represent, respectively, an enzyme and a compound. The enzymes whose genes are identified in the genome are shown by colored (shaded) rectangles.

have only the lower and upper half of the TCA cycle, respectively, although a missing enzyme in *H.pylori* still needs be identified to make the pathway continuous.

The PATHWAY database can be retrieved most conveniently from the top menu of the KEGG table of contents page, metabolic pathways and regulatory pathways that are categorized by the

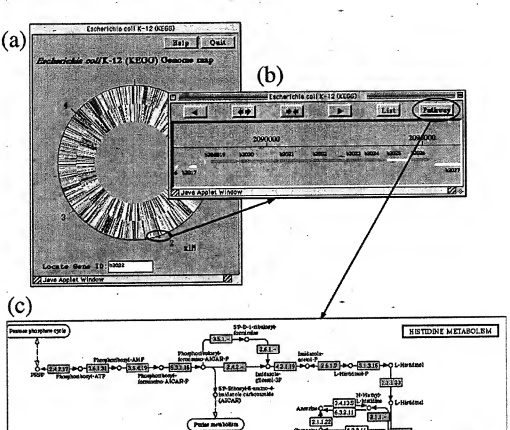

Figure 2. The correlation of a physical unit and a functional unit of genes. (**a**) Starting from the KEGG genome map for *Escherichia coli*, (**b**) a chromosomal region is selected in the zoom-up window. (**c**) By clicking on the 'Pathway' button, the user can examine if the biochemical pathway can be formed from a set of genes in the zoom-up window.

hierarchical classification of Table 2. Alternatively, the PATHWAY database can be searched by the DBGET/LinkDB system.

GENES database and gene catalogs

The KEGG/GENES database is a collection of genes for all organisms in KEGG that is organized as a flat-file database of textual information. In fact, each organism is a primary database and GENES is defined as a composite database of all organisms under the DBGET/LinkDB system. An entry of the GENES database contains the information: organism name, gene name, functional description, functional hierarchy (KEGG pathway classification), chromosomal position, codon usage, amino acid sequence, and nucleotide sequence (Table 1). The genes in each organism are hierarchically classified in the gene catalog according to the KEGG pathways (Table 2) and they can be viewed and searched as, what we call, hierarchical texts in KEGG. The entire GENES database or each organism separately can also be searched by the DBGET/LinkDB system.

The GENES database is maintained as follows. First the information of all genes in an organism is automatically generated from the complete genomes section of the GenBank database (10). Then the EC number assignment is performed by GFIT (8) and other programs with manual verification efforts. The gene function annotations are continuously re-evaluated according to the KEGG/PATHWAY database and by comparing with SWISS-PROT (12) and other databases.

Genome maps and comparative genome maps

The genome map represents a one-dimensional network of genes that are physically located in a circular or linear genome. The gene order turns out to be extremely valuable information in functional annotation, especially for bacteria and archaea. As shown in Figure 2, the genome map is linked to the pathway map by the Pathway button. The user can check if a physical unit of closely located genes (e.g., genes in an operon) would form a functional unit of related proteins that appear at close positions in

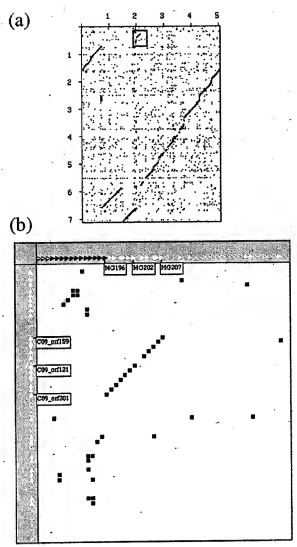

Figure 3. The comparative genome map in KEGG. (**a**) An overview of comparing *Mycoplasma genitalium* (horizontal) and *Mycoplasma pneumoniae* (vertical) genomes is shown where each dot represents significant amino acid sequence similarity of two genes. (**b**) The zoom-up window for the boxed area in (**a**) is shown where each gene name can be identified and pre-calculated homologous gene strings can optionally be displayed in color.

the pathway. By clicking on the List button in the genome map, the user can also invoke a sequence similarity search to see if a stretch of genes would match a functional unit in the pathway.

The co-linearity of genes between two genomes is quite useful for identification of clusters of orthologous genes. KEGG provides the comparative genome map for identification of such clusters and for functional annotation of newly sequenced genomes (Fig. 3). The comparative genome map is analogous to the dot matrix for comparing two nucleotide or amino acid sequences. Here, a dot represents significant sequence similarity between two genes at the amino acid sequence level. To present this map, the sequence similarity scores are pre-computed by SSEARCH for all pairs of organisms and homologous gene clusters are pre-identified so that they can be optionally highlighted in the map.

Ortholog group tables

The ortholog group table is a summary table that represents functional correlations in the pathway, physical (positional) correlations in the genome, and evolutionary (sequence similarity) correlations among species. Currently there are about 4800 genes that belong to about 140 groups of functional units. The ortholog group table is most useful as a reference data set for functional annotations. KEGG provides a computational tool to search against the ortholog group tables for sequence similarity of a set of query sequences (see below).

A case in point is the annotation of ABC transporters. A typical ABC transporter consists of a substrate-binding protein, two membrane proteins, and two ATP-binding proteins. The genes for these molecular components usually form an operon, and there are many paralogous genes that are responsible for the transport of different substances. If a set of query sequences that are located at physically close positions in the genome match against all components, it is a good indication that the transporter unit is correctly reconstructed. Since the KEGG ortholog group table for ABC transporters is functionally categorized (9), the pattern of matches and their similarity scores can be used to deduce substrate specificity.

Molecular catalogs and LIGAND database

The KEGG molecular catalogs are intended to provide functional and structural classifications of proteins, RNAs, other biological macromolecules, chemical substances and molecular assemblies. However, the organization is still rudimentary except for enzymes. The information of chemical compounds, enzyme molecules, and enzymatic and non-enzymatic reactions is stored in the LIGAND database as described in the accompanying paper (5). Based on this database KEGG provides several molecular catalogs for classifications of enzymes and a preliminary classification of chemical compounds.

FUNCTIONAL GENOMICS

Pathway reconstruction with reference maps

The KEGG reference maps for metabolic pathways represent biochemical knowledge containing all chemically identified reaction pathways. The constraint of the genome, i.e., a list of enzymes encoded in the genome, will reconstruct organism-specific pathways, which are represented by coloring of boxes in the KEGG pathway maps (Fig. 1). Furthermore, the constraint of operons will often predict functional units or conserved pathway motifs, which are represented by additional coloring in the KEGG pathway maps (Fig. 2). The operon information probably reflects a regulatory unit of transcription. Thus, it is easy to see how the information of gene expression profiles can be used as still another constraint against the KEGG reference pathway maps. In fact, KEGG provides a tool to color the pathway maps in order to visualize, for example, the microarray patterns of gene expression profiles.

Table 3 shows the list of currently available tools for search and analysis of KEGG pathway maps and genome maps. The user-interfaces for these tools can be accessed from the KEGG

Table 3. Computational tools in KEGG

Category	Tools
Search tools	• Search pathway maps by gene names, EC numbers, and compound numbers • Search genome maps by gene names • Hierarchical text search • DBGET/LinkDB search
Coloring tools	• Color genes in the pathway map as specified, for example, by the microarray patterns • Color genes in the genome map as specified
Prediction tools	• Reconstruct pathways from a set of genes by sequence similarity search against pathway maps • Reconstruct pathways from a set of genes by sequence similarity search against ortholog group tables • Generate possible reaction pathways between two compounds
Gene cluster search	• Search homologous gene clusters in genome maps • Compare two genome maps for exhaustive search of homologous gene clusters
Sequence similarity search	• Search similar amino acid sequences in the GENES database by FASTA or BLAST • Search similar nucleotide sequences in the whole GENOME database by FASTA or BLAST

table of contents page. The pathway reconstruction tools in the category of prediction tools are based on sequence similarity search that involves a set of query sequences at a time to see if a pathway is correctly reconstructed. At the moment, there are two versions of the reconstruction tools. One is to search against the KEGG pathway maps: http://www.genome.ad.jp/kegg-bin/mk_homology_pathway_html , and the other is to search against the ortholog group tables: http://www.genome.ad.jp/kegg-bin/srch_orth_html . The former contains a larger data set of pathways but the search is made against single organisms. The latter is limited to selected portions of the pathways (pathway motifs or functional units), but the search is made against multiple alignments of organisms and tends to produce better results.

Pathway reconstruction from binary relations

It is often the case that no homology can be found when searching against the KEGG reference pathways, which suggests that the current biochemical knowledge is not sufficient to predict a pathway. If there is a missing portion in the reconstructed metabolic pathway, KEGG provides a tool to predict alternative paths with alternative enzymes. The tool actually computes all possible reaction pathways between two compounds from a set of substrate–product relations, i.e., from a set of enzymes (5): http://www.genome.ad.jp/kegg-bin/mk_pathcomp_html

This tool also has a feature called query relaxation to incorporate grouping or hierarchy of relations. Whenever any member of the group is identified in the genome by sequence similarity (e.g., an enzyme in the same hierarchy of EC numbers), the entire group is incorporated for computation (e.g., to represent wider substrate specificity). This effectively increases the number of substrate–product relations and expands possible reaction pathways.

This type of computation, which we call pathway reconstruction from binary relations, can be performed in a more general way. A binary relation can be a substrate–product relation in metabolic pathways, a gene–gene interaction observed in gene expression profiles, or a protein–protein interaction observed by yeast two hybrid system experiments. In practice, the reconstruction from binary relations works better when the problem size is not large. For example, if most of the pathway is reconstructed from the reference but there still remain a few missing enzymes, then the computation with binary relations may fill the missing links of fragmented pathways. In addition to the substrate–product binary relations, KEGG will provide tools to integrate and compute different types of binary relations. Perhaps this is the most challenging area of computational problems that have become accessible by the KEGG project.

AVAILABILITY

The Internet version of KEGG can be accessed at the following address:

http://www.genome.ad.jp/kegg/

For strictly academic purposes at academic institutions the KEGG mirror server package may be installed. The package, which also includes a minimal set of DBGET/LinkDB, can be obtained from the KEGG anonymous FTP site:

ftp://kegg.genome.ad.jp/

The mirror package runs on a Solaris or IRIX machine. The individual databases PATHWAY, GENES, and LIGAND can also be obtained from this FTP site.

The CD version of KEGG was once distributed and a copy still exists at the FTP site. However, since the database has become so large and since the system is undergoing frequent revisions, the CD version is not supported at the moment. We will redefine the role of CD and hope to start distributing again. Finally, some of the search tools are also available at the KEGG mail server:

mailto:pathway@genome.ad.jp

ACKNOWLEDGEMENTS

This work was supported by the Grant-in-Aid for Scientific Research on the Priority Area 'Genome Science' from the Ministry of Education, Science, Sports and Culture of Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

REFERENCES

1 Kanehisa,M. (1997) *Trends Genet.*, **13**, 375–376.
2 Goto,S., Bono,H., Ogata,H., Fujibuchi,W., Nishioka,T., Sato,K. and Kanehisa,M. (1997) *Pacific Symp. Biocomput. 1997*, 175–186.
3 Ogata,H., Goto,S., Fujibuchi,W. and Kanehisa,M. (1998) *BioSystems*, 47, 119–128.
4 Goto,S., Nishioka,T. and Kanehisa,M. (1998) *Bioinformatics*, 14, 591–599.
5 Goto,S., Nishioka,T. and Kanehisa,M. (1999) *Nucleic Acids Res.*, 27, 377–379.
6 Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa,M. (1998) *Pacific Symp. Biocomput. 1998*, 683–694.
7 Kanehisa,M. (1997) *Trends Biochem. Sci.*, **22**, 442–444.
8 Bono,H., Ogata,H., Goto,S. and Kanehisa,M. (1998) *Genome Res.*, 8, 203–210.
9 Tomii,K. and Kanehisa,M. (1998) *Genome Res.*, 8, 1048–1049.
10 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) *Nucleic Acids Res.*, **26**, 1–7.
11 Ermolaeva,O., Rastogi,M., Pruitt,K.D., Schuler,G.D., Bittner,M.L., Chen,Y., Simon,R., Meltzer,P., Trent,J.M. and Boguski,M.S. (1998) *Nature Genet.*, **20**, 19–23.
12 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42